DOCUMENT RESUME

ED 144 979 TM 006 529

AUTHOR Crawford, John; And Others
TITLE Methods for Maximizing the Validity of Experiments on Teaching.
PUB DATE Apr 77
NOTE 25p.; Paper presented at the Annual Meeting of the American Educational Research Association (61st, New York, New York, April 4-8, 1977)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Classroom Research; Control Groups; Data Analysis; *Educational Experiments; Experimental Groups; *Field Studies; *Research Design; *Research Problems; Teacher Behavior; *Teaching; *Validity

ABSTRACT

A number of potential threats to internal validity in true experiments are discussed. Programmatic research efforts proceeding from established correlational findings to experiments employing random assignment are described and used as referents in discussions of validity considerations. The collection of observational data is recommended for use in analyses of the degree of treatment implementation. The summary presents examples of the near-inevitable tradeoff between different types of validity (internal and external) in experiments on teaching. (Author)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED144979

Methods for Maximizing the Validity of Experiments on Teaching

John Crawford

N. L. Gage

Jane A. Stallings

Stanford University

School of Education

Stanford Center for Research and Development in Teaching

Program on Teaching Effectiveness

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

John Crawford

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

TM006 529

(Session 27.11)

This paper was presented as part of a symposium on "Rigor and Relevance in Experiments on Teaching in Classrooms" at the annual meeting of the American Educational Research Association, New York, N. Y., April 1977.

Abstract

A number of potential threats to internal validity in true experiments are discussed. Programmatic research efforts proceeding from established correlational findings to experiments employing random assignment are described and used as referents in discussions of validity considerations. The collection of observational data is recommended for use in analyses of the degree of treatment implementation. The summary presents examples of the near-inevitable tradeoff between different types of validity (internal and external) in experiments on teaching.

Methods for Maximizing the

Validity of Experiments on Teaching

In recent years, several correlational studies have examined process-product relationships in elementary-level classrooms. These studies, by Brophy and Evertson (1974), McDonald and Elias (1976), Soar (1973), and Stallings and Kaskowitz (1974), have suggested a number of teaching practices which are predictive of adjusted pupil achievement. The teacher training component of the experiment being conducted by the Program on Teaching Effectiveness (in press) at Stanford was designed to test the causal nature of relationships between teaching behaviors, as measured in the studies cited above, and pupil achievement.

Such an experiment was judged desirable in view of the relatively high degree of consistency of findings in the four major process-product correlational studies. The data from these studies reflect only the natural covariation of teaching behavior and pupil achievement, and hence allow no valid causal inferences concerning whether the variation in the selected teaching behaviors was actually causing the varying levels of pupil achievement. Before findings from the correlational studies become directly incorporated into preservice teacher education programs, experimental verification of the causal relationships between teaching practices and pupil measures should be available.

An experimental test of the causal efficacy of results from the observational studies is the next step in what Rosenshine and Furst (1973) have described as the "descriptive-correlational-experimental" loop. That

is, programmatic research may be most productive when it first seeks an adequate description of the phenomena of interest, then establishes a complex of relationships suggestive of the nature of the covariation between the classroom process measures and outcome variables, and finally proceeds to the design and conduct of true experiments which employ random assignment to treatment groups and systematic intervention with an experimenter-developed treatment.

An earlier experiment by the Program on Teaching Effectiveness (1976) exemplifies the progression from correlates of pupil achievement and attitude to a field-based experimental investigation of major variables suggested by the literature. In that experiment, teacher structuring, soliciting, and reacting were manipulated. Where evidence from such experiments does support the hypotheses suggested by correlational studies, it becomes possible to have confidence in both the generalizability of the correlational results and the direction of causality of the relationships--such that the manipulated teaching behaviors can be said to be causing the pupil growth being measured as the dependent variable.

In a recent follow-up and expansion on issues raised in Campbell and Stanley (1963), Cook and Campbell (1976) have discussed a number of considerations relating to the validity of experimentation in field settings. The Cook-Campbell chapter explicates four kinds of validity: internal, construct, statistical conclusion, and external. By internal validity, Cook and Campbell mean the degree to which an experiment demonstrates that a causal relationship from A to B is, in fact, not mediated or caused by some third variable or set of variables. Construct validity refers to the correct labeling of the cause and effect; for example, a "treatment effect"

might in some instances be due to the overall amount of attention given to subjects in the experimental condition, rather than to the actual treatment employed (in which case the results would have to be attributed to the Hawthorne effect). Statistical conclusion validity relates to the possibility of false conclusions--about the relationship between treatments and effects--on the basis of inferential statistical tests. External validity is the degree to which the findings may be generalized across time, settings, and persons.

Cook and Campbell also present three necessary conditions for causal inferences:

(1) temporal antecedence of cause (before effect)

(2) covariance of treatments with the observed effects

(3) absence of a plausible explanation of the observed effects other than the manipulated treatments.

Since the third condition is the most difficult to meet, and since social science methodologists (Campbell & Stanley, 1963; Cook & Campbell, 1976) generally agree that internal validity is the sine qua non of causal inference, this paper will concentrate on threats to internal validity and ways of dealing with these threats. One section briefly discusses factors relating to external, construct, and statistical conclusion validity. The discussion of methods for dealing with threats to validity will refer for the most part directly to the experiment now being conducted by the Program on Teaching Effectiveness (PTE) (see PTE, in press). This will allow the discussion to be grounded in a research effort being wholly conducted in regular public school classrooms.

The Experiment on Teacher Effectiveness and Parent-assisted Instruction in the Third Grade

Since the remainder of the paper will refer to the experiment now being conducted by PTE, this section will describe briefly the design and procedures of that experiment.

The detailed examination of correlational findings (from the four studies cited) in the second and third grades made possible the development of an experimental teacher training program with recommendations for teachers in three content areas: behavior management, instructional methods, and questioning and feedback strategies. Even after an extensive synthesis and reduction process, the number of distinct recommendations for teachers remained substantial (more than 20). So a decision was made to employ a simple "training vs. no training" design, rather than to attempt manipulating each behavior independently--as in a complex factorial or a fractional factorial design.

The final design (schematically represented in Figure 1) used three levels for the teacher training factor, crossed with two levels of parent involvement (yes vs. no) in the Learning Skills Program. The parent-assisted Learning Skills Program will not be referred to here, since the present discussion is concerned with the validity of experiments on teaching.

Insert Figure 1 about here

The three levels of the teacher training factor have been labeled "minimal," "maximal," and "observed-only" (control) treatments. The content of the training given to teachers in the minimal and maximal groups was the same, but the mode of presentation varied. The minimally trained teachers

received written packets only, while the maximally trained teachers received the same written packets and attended five weekly two-hour meetings to discuss and practice the teaching strategies with PTE staff. The design was replicated in two school districts.

A stratified random assignment procedure was used in placing classrooms in the cells defined by the teacher training and parent involvement factors. Classes were first stratified according to pupils' pretest scores on the reading achievement tests (to be used as a major dependent variable) and were then randomly assigned to the treatment conditions so that each cell contained classes from each of the different achievement strata.

In addition to manipulations associated with the teacher training and parent involvement factors, other experimental activities included pretesting (in early Fall 1976) of pupils and teachers with a variety of cognitive and attitudinal instruments, mid-year testing of pupils and interviews with teachers (during January and February 1977), and posttesting of pupils and teachers (scheduled for May 1977). A large amount of low- and high-inference observational data will be used in determining the implementation of the teacher training treatments; low inference data are being collected with a modified version of the SRI Classroom Observation Instrument (Stallings & Kaskowitz, 1974). These data consist of four full days of observation in all classrooms prior to the experimental treatments (baseline observational data), five full days of observation during the five week treatment period (treatment data), and seven to eight days of observation after the treatment period (post-treatment data). As will be noted later, these data provide the major means for minimizing a number of threats to the validity of the experiment.

Problems and Procedures Related to Internal Validity

Much of the chapter by Cook and Campbell (1976) is devoted to quasi-experimental designs and ways of maintaining acceptable levels of validity in these types of studies. Here we concentrate on problems associated with true experiments since the study now being conducted by the Program on Teaching Effectiveness is of this type. This permits a discussion more closely focused on true field-based experiments on teaching. Cook and Campbell noted that the use of true rather than quasi-experimental designs (where true experiments are those employing random assignment and treatments developed and administered by the experimenter) eliminates many of the major threats to internal validity by forming initially similar groups, providing similar testing conditions for all subjects, and generally equating the historical experiences of the different groups during the experiment.

Yet, a number of threats to internal validity remain--threats not eliminated simply by conducting a true experiment. The remainder of this section deals with the following factors: diffusion of treatment, compensatory equalization, compensatory rivalry, resentful demoralization, and local history effects.

Diffusion of Treatment

The validity threat which Cook and Campbell call "diffusion or imitation of the treatment" refers to the possibility that experimental and control groups may be able to communicate with each other, thereby increasing the likelihood that some subjects assigned to the control group condition would obtain knowledge of the experimental treatment. This is particularly problematic when more than one teacher at the same school is participating in the experiment, and these teachers are not in the same treatment group, and, to a lesser extent, when the experiment is being conducted in

a school district where several schools are participating and teachers in different schools but the same district are assigned to different groups. Face-to-face communications seems assured between teachers in the same school, particularly if they are teaching at the same grade level.

One method of minimizing this threat would be to study no more than one teacher in each school. To achieve larger Ns per cell, the PTE experiment used some schools with more than one teacher--the 33 teachers are in 24 schools, so there are nine instances of two teachers in the same school. The random assignment of teachers to the treatment groups yielded three instances (schools) in which the "worst" match-up occurred: where a control group (observed-only) teacher was in the same school with a teacher receiving one of the two training treatments (minimal or maximal). In the other six schools where more than one teacher was participating, both teachers were either in the control (observed-only) group, or one was assigned to a treatment group receiving one mode of training, and the other was assigned to a group receiving the other mode (but the same content).

Steps were taken initially to explain to teachers (before they agreed to participate) that this was an experiment in which we would be asking them to refrain from communicating about the experimental treatments. Also, written reminders were sent with the experimental packets (the teacher training materials) asking the teachers not to communicate with others concerning the experimental recommendations.

Nonetheless, it was still considered possible that some teacher communication about the experimental materials would occur. Because of this, an interview was designed to get (along with other information) the

teachers' own reports of the degree to which they communicated with others about the experimental treatment. Preliminary analyses of those interview data have indicated that, while some communication did occur, it was relatively limited in scope, and restricted to a small portion of the total sample of teachers.

It should also be possible to examine the observational data to determine whether teachers in the control groups in schools where there was also a trained teacher evidenced behavior that was more congruent with the training recommendations than control-group teachers in schools where no training group teacher was present. Since the training recommendations for teachers were derived from observational studies of natural variation, it is certainly possible that the control (observed-only) teachers would be using some of the experimental techniques without having been told about them. Our hope is that the trained and observed-only teachers will differ substantially in the rate and consistency of their use of the training recommendations, in a direction favoring the trained teachers.

Compensatory Equalization of Treatment

This threat to internal validity refers to the perception, usually on the part of administrators, that the distribution of experimental programs to some subjects but not others should be "compensated" by giving some alternative program to the control group not scheduled to receive any intervention. The example given by Cook and Campbell referred to the Follow Through Program administration: the "control group" schools not receiving Follow Through funds were often given Title 1 funds in amounts equivalent to those going to the Follow Through schools. This threat should not be a problem in the experiment now being conducted by PTE since

the complete experimental design was explained to district personnel before the experiment was begun--with the stipulation being accepted by the administrators that some teachers would not receive any experimental treatments. The district officials and school principals were told that any teachers who were randomly assigned to the control group could receive the teacher training program following the completion of the experiment. The explanation given was that, unless we could withhold the experimental treatments from some teachers between pre- and posttesting it would not be possible to determine whether a causal relationship existed between our training program and the intended outcome (i.e., improved pupil reading achievement).

Compensatory Rivalry

The compensatory rivalry factor is similar in operation to compensatory equalization of treatment, except that the compensation is being instigated by control group subjects, rather than administrators or "higher level" personnel with the power to equalize treatments. When a control group receiving no alternative or placebo treatment is used, and all the subjects, including the control group, are given a general outline of the experiment, this factor may pose problems. The control group may perceive itself as being the "underdog" and increase efforts to show high performance on the dimension perceived as the intended outcome of the experiment.

Since all teachers participating in the present experiment know that there are some trained and some untrained teachers, and that the primary outcome being studied is pupil achievement, it would be relatively easy for control-group teachers to spend more time on reading instruction and basic

academic skills--which would tend to increase their pupils' scores on standardized reading achievement tests. Saretsky (1972) labeled this kind of effort by a control group the "John Henry" effect, referring to the mythical character who outperformed a competing machine, but died of exhaustion in the end. While there is not really any way to control for this effect ahead of time (particularly when the control group is not receiving any placebo treatment), it is possible to determine through ex post facto analyses the degree to which such efforts may have occurred. These analyses would be based on the implementation data, which would indicate whether the teacher worked particularly diligently with his or her students in areas likely to be related to the major outcome variable (pupil reading achievement).

These three internal validity threats--diffusion of treatment, compensatory equalization, and compensatory rivalry--would probably tend to equalize the scores of experimental and control groups on dependent variables. We turn now to other factors--"resentful demoralization" and "local history"--that might tend to create a spurious difference between experimental and control groups.

Resentful Demoralization of Respondents
Receiving Less Desirable Treatments

Cook and Campbell aptly called the control group in most obtrusive experiments the "resentful-of-no-treatment" group. By this designation, they are indicating that the control group participants may have developed an attitude that would tend to decrease their performance on the dependent measure. This would increase the likelihood of obtaining a significant difference in "treatment vs. control" comparisons. This effect might be

counterbalanced somewhat by the "John Henry" effect mentioned under compensatory rivalry. In the present PTE experiment this phenomenon may not be restricted to the control group, since the minimally trained teachers are receiving a treatment which could be considered less desirable than the treatment received by maximally trained teachers. Although this factor cannot really be controlled, its impact might be minimized by de-emphasizing the comparisons to be made across treatment groups and perhaps by suggesting that the treatment is actually an "experimental try-out" of a new program that may or may not be effective.

Local History

Random assignment to treatment groups should preclude any major historical differences between the experiences of different groups during the experiment. But there may be reactive effects of the treatment, also entailing some "local history" effects, which would vary with assignment to experimental groups. Events that are probably irrelevant (but may be relevant) might not occur equally for the experimental and control groups. For example, the assignment of all experimental participants to one group session would allow the possibility that fortuitous occurrences before, during, or after that session could have effects related to the experimental treatment, and hence, the dependent measure score. To some extent, our replication of the experiment in two school districts guards against this threat; it is unlikely that the same fortuitous bit of "history" would occur in both districts.

Cook and Campbell noted that these kinds of local history effects are more likely to be evident in quasi-experiments than in true experiments. In the present PTE study, this should not pose a great problem. One of the

participating districts did have a teachers' strike during the period preceding administration of the experimental treatment. However, both the control and the experimental teachers in the district involved with the strike were participating in equal numbers, and therefore there should not be any confounding of the strike with assignment to treatment group.

Considerations Relating to Construct, Statistical Conclusion, and External Validity

Under the heading of construct validity, Cook and Campbell discuss the selection "of measures or manipulations that represent the [intended] abstract construct" (p. 239). Construct validity refers primarily to the degree to which the cause of the measured effect was labeled appropriately. The so-called Hawthorne effect is a good example of how an effect might be inappropriately attributed to the manipulated treatment when, in fact, the effect might be due simply to the overall amount of attention given to the experimental group(s).

One way of dealing with this problem is to administer a placebo treatment to the control group, such that the treatment given to these subjects requires about the same amount of time and attention as the experimental treatment, but is not related to the dimensions that are actually of interest and are being investigated as the experimental treatment. Such a placebo treatment was not used in the present experiment because of a desire to maintain the representativeness of the situation in which an inservice program would typically be implemented and evaluated. That is, we wanted a comparison of the inservice program with the status quo (no program) rather than a comparison of one inservice program with another unrelated one.

A relatively high level of construct validity was sought through careful development of a conceptual framework for the findings of the four correlational studies cited earlier and through relating those findings to meaningful components of classroom instruction. So, while construct validity was not maximized methodologically by employing a placebo treatment with the control group, conceptual steps were taken to maximize the manipulation of the intended constructs (i.e., those derived from the four correlational studies). The teacher training packets that served as the experimental treatment were developed by a process whereby operational definitions, means, standard deviations, metrics of process variables, and authors' interpretations of results were examined in an attempt to explicate the constructs prior to developing the experimental treatments. See Crawford and Gage (in press) and PTE (in press) for a more detailed description of the process of developing the teacher training recommendations.

A high level of construct validity may also be maintained by measuring dependent variables manifestly pertinent to the intended dimensions and by having good, reliable measures of the "take" of the experimental treatment. These conditions were met in the present PTE experiment by using multiple dependent variables--achievement data are coming from program-administered test batteries and from the school district data, and by collecting a large amount of observational data on classroom activities and teacher and pupil behavior. The observational data will allow a determination of how closely and consistently the teachers in the experimental groups were able to follow the recommendations about ways of teaching. Several tests of reading achievement and inventories

measuring other outcomes will permit examination of the consistency of findings across a number of measures.

Threats to the validity of statistical conclusions present the greatest potential problems affecting inferences from the present experiment. Since the experiment is being conducted in public school settings with the regular classroom teachers, there is a problem of "reliability" or heterogeneity of treatment implementation. Cook and Campbell note that whenever a treatment is implemented somewhat inconsistently across subjects in a treatment group, error variance is increased. Accordingly, the chance of obtaining significant differences between treatment groups tends to decrease. However, if treatment group effects are discovered in spite of such problems in the reliability of treatment implementation, one may assume that such effects are more real than spurious.

Also, the present experiment only manipulated the mode of teacher training as an independent variable. The content of the training was the same for the teachers in the minimal and maximal training groups. The difference was in the delivery system--the maximally trained teachers came to personal, interactive sessions with PTE staff, whereas the minimally-trained teachers received and completed the written materials at home on their own. Therefore, the discrete variables which resulted from the analysis of the correlational literature were not manipulated independently. To determine the effects of individual variables, it will be necessary to use a combination of analyses. Such analyses include multiple regression, process-product rs, path analyses, and other multivariate methods.

The implementation analyses will also be useful in forming conclusions about individual variables, since an overall treatment effect would not be attributable to any discrete process variable not found to discriminate

between untrained and experimental groups of teachers. While no single statistical test may demonstrate the effects of discrete process variables, analyses of the data from a number of different perspectives may yield a convincing picture of causal relationships between specific teaching practices and measures of pupil achievement and attitude.

While field-based experiments do not automatically have high external validity, the present study seems to represent "real-world" conditions as well as possible while still employing sufficient internal controls. The representativeness resides in the use of regular classrooms, teachers, and their pupils in two different public school districts. Also, the experimental treatments are being administered in standard inservice teacher education formats, one in a "training workshop" format and the other in a "read-only" format. While the content of the classrooms' curricula is not being controlled (as it would with a programmed instruction task or with especially trained substitute teachers presenting a scripted lesson), the greater representativeness of presenting the teacher training program to the regular classroom teachers should compensate for any decrement in internal validity controls.

The major factors affecting the representativeness of the sample include the location of the two school districts being used, and the fact that all participating teachers were volunteers. The relative achievement and SES level of the pupils in the two districts near Stanford will be ascertainable in terms of State norms. Also, participating teachers are probably not representative of the larger population of teachers, since they all volunteered for participation in the experiment. While this

should not affect comparisons across treatment groups, it does have implications for any subsequent mandated implementation of such a teacher training program, since the present teachers, as compared with nonvolunteers, are more highly motivated to improve their teaching, more secure, and less worried by activities such as having an observer in their classroom on many days for the entire year. Access to statewide data will permit us to establish empirically the representativeness of the present sample on certain dimensions.

Final Comments

While a high level of one kind of validity may not necessarily imply a low level of another kind of validity, such tradeoffs are frequent enough to appear inevitable. The most tightly controlled studies in experimental psychology often result in findings which, strictly speaking, should not be generalized beyond college sophomores working in a contrived laboratory setting. But the same studies are the ones which may result in the most unambiguous inferences that treatment manipulation "A" causes outcome "B". Likewise, field-based experiments in teaching may have a high degree of generalizability to related populations and situations, but may yield equivocal findings that leave some doubt as to whether "A" actually caused "B".

Following Cook and Campbell, we would suggest that experimenters attach priorities to the factors affecting validity and deal with the most crucial aspects of experimental design before moving to lesser concerns. In general, internal validity should probably have the highest priority, although this may not always be the case when true experiments are being conducted and major threats to internal validity have been removed simply

by the use of randomization and experimenter-developed treatments. The ranking of external validity in a hierarchy of validity concerns would depend primarily on the researchers' goals and aims and whether the research is intended to be "basic" or "applied."

A comparison of the two most recent experiments conducted by PTE (the more recent being the one presently in progress) illustrates the tradeoff between more tightly controlled and more generalizable experiments. The former study was an orthogonal manipulation of three teaching behaviors (teacher structuring, soliciting, and reacting) in a factorial design.

Experienced teachers were hired by the Program and extensively trained in methods of teaching that could be classified as "high" or "low" on each of the three independent variables. These teachers then went to selected classrooms and administered the treatments by teaching a series of nine scripted 40-minute lessons over a two-week period. All students were taught the same content, so that only the instructional dimensions of structuring, soliciting, and reacting varied. Implementation data indicated that the teachers did accurately render the experimental treatments and were well able to follow the experimental scripts.

That study may be seen as having a high degree of internal validity but lacking in demonstrated generalizability to a larger population of teachers and contexts--since the four teachers used in the experiment were highly trained and were not allowed to deviate from the content of the lessons that formed the experimental treatments. Further, since unexpected differences in achievement associated with teachers did occur, we must note that the use of trained experimental teachers, rather than the regular classroom teachers, does not insure elimination of all potentially confounding effects.

In contrast, the present study being conducted by PTE is using regular classroom teachers for implementing the experimental treatments and is applying no control over the content of the curriculum being taught. While the experimental manipulations may be implemented with greater variability and the curriculum content may, in some way, be reactively affecting the posttest measures, this experiment does have a much higher degree of external validity. The treatment was administered as a standard inservice teacher education program, and the normal ecology of the classroom was disturbed very little by the experimental intervention.

While there are fewer controls in the present study, the collection of a large amount of observational data makes it possible to determine the extent of various possible confounding effects--such as inconsistent implementation of the treatment recommendations. The replication of the design in two school districts also allows a broader test of generalizability of results. If effects are evident in one district but not the other, demographic and other data may be examined in an attempt to determine the cause of such lack of replication.

While true experiments do not necessarily have an advantage over quasi-experiments with respect to external validity, they do typically have higher levels of internal validity (and statistical conclusion validity). The use of a random assignment procedure seems to be advisable in every possible instance if the researchers wish to make strong statements about causal relationships between manipulated independent variables and outcome measures.

In initial meetings with teachers, the PTE staff did meet with some sentiment such as "Why can't we volunteer for the group we would like to be in?" The response to this inquiry was that, if we allowed them to

volunteer for the group that they wanted to be in, we would have no way of knowing, when the experiment was over, whether any obtained effects were due to the intervention in classrooms by PTE, or whether they were due to the initial volunteering of the teachers into the different groups. The teachers indicated that this sounded reasonable and all of those who then agreed to participate also agreed to be placed in whichever group they were randomly assigned to. This approach agrees with the findings of Hillis and Wortman (in press) that public acceptance of randomized experimental designs can be best facilitated by careful explanation of the scientific necessity for randomization. Without going into a jargon-laden explanation of internal validity and causal efficacy, it is certainly possible to convey the idea that random assignment is the only way to insure that treatment groups are initially equivalent.

With a certain amount of care and planning, field-based experiments may prove to be promising approaches to the investigation of social phenomena. Without experimentation, the causal nature of important relationships will remain ambiguous, and advancement beyond statements of covariation will not be possible.

References

Brophy, J., & Evertson, C. Process-product correlations in Texas Teacher Effectiveness Study: Final Report. Research Report No. 74-4. Austin, Texas: Research and Development Center for Teacher Education, University of Texas, 1974.

Campbell, T., & Stanley, J. Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), Handbook of research on teaching. Chicago: Rand McNally, 1963. (Also published as Experimental and quasi-experimental designs for research. Chicago: Rand McNally, 1966.)

Cook, T., & Campbell, D. The design and conduct of quasi-experiments and true experiments in field settings. In M. Dunnette (Ed.), Handbook of industrial and organizational psychology. Chicago: Rand McNally, 1976.

Crawford, J., & Gage, N. Development of a research-based teacher training program. California Journal of Teacher Education, in press.

Hillis, J., & Wortman, C. Some determinants of public acceptance of randomized control group experimental designs. Sociometry, in press.

McDonald, F., & Elias, P. The effects of teaching performance on pupil learning, Beginning Teacher Evaluation Study: Phase II, Final Report: Vol. I. Princeton, N.J.: Educational Testing Service, 1976.

Program on Teaching Effectiveness, SCRDT. A factorially designed experiment on teacher structuring, soliciting, and reacting (R&D Memorandum No. 147). Stanford, Ca.: Stanford Center for Research and Development in Teaching, 1976.

Program on Teaching Effectiveness, CERAS. An experiment on teacher effectiveness and parent-assisted instruction in the third grade. Stanford, Ca.: Center for Educational Research at Stanford, in press.

Rosenshine, B., & Furst, N. The use of direct observation to st dy teaching. In R. M. W. Travers (Ed.), Second handbook of research on teaching. Chicago: Rand McNally, 1973.

Saretsky, G. The OEO P.C. experiment and the John Henry effect. Phi Delta Kappan, 1972, 53, 579-581.

Soar, R. Final report, Follow Through classroom process measurement and pupil growth (1970-1971). Gainesville, Fla.: Institute for the Development of Human Resources, University of Florida, 1973.

Stallings, J., & Kaskowitz, D. Follow Through classroom observation evaluation, 1972-1973. Menlo Park, Ca.: Stanford Research Institute, 1974. Office of Education Contract OEC-0-8522480-4633(100).

Teacher Training

District I[a]		Maximal	Minimal	Control	Totals
Parent Involvement in the LSP[b]	Yes	3	3	3	9
	No	3	3	3	9
	Totals	6	6	6	18

District II

Parent Involvement in the LSP	Yes	3	3	2	8
	No	3	2	2	7
	Totals	6	5	4	15

[a]The design was replicated in two different elementary school districts.

[b]LSP: Learning Skills Program

Numbers represent the classroom Ns in each treatment condition

Figure 1. Schematic representation of the experimental design, including treatment group Ns